IJDC | *Conference Paper*

Cluster Analysis of Open Research Data: A Case for Replication Metadata

Ana Trisovic
Institute for Quantitative Social Science
Harvard University
Cambridge, MA, USA
anatrisovic@g.harvard.edu

Abstract

Research data are often released upon journal publication to enable result verification and reproducibility. For that reason, research dissemination infrastructures typically support diverse datasets coming from numerous disciplines, from tabular data and program code to audio-visual files. Metadata, or data about data, is critical to making research outputs adequately documented and FAIR. Aiming to contribute to the discussions on the development of metadata for research outputs, I conduct an exploratory analysis to determine how research datasets cluster based on what researchers organically deposit together. The content of over 40,000 datasets from the Harvard Dataverse research data repository is used as a sample for the cluster analysis. I find that the majority of the clusters are formed by single-type datasets, while in the rest of the sample no meaningful clusters can be identified. For the result interpretation, I use the metadata standard employed by DataCite, a leading organization for documenting a scholarly record, and map existing resource types to my results. About 65% of the sample can be described with a single-type metadata (such as *Dataset*, *Software* or *Report*), while the rest would require aggregate metadata types. Though DataCite supports an aggregate type such as a *Collection*, I argue that a significant number of datasets, in particular those containing both data and code files (about 20% of the sample), would be more accurately described as a *Replication resource* metadata type. Such resource type would be particularly useful in facilitating research reproducibility.

Submitted date August 2022 ~ *Accepted* date January 2023

Correspondence should be addressed to Ana Trisovic, 1737 Cambridge St, Cambridge, MA 02138. Email: anatrisovic@g.harvard.edu

This paper was presented at International Digital Curation Conference IDCC22, online, 13-16 June, 2022

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v17i1.833
DOI: 10.2218/ijdc.v17i1.833

Introduction

Computational research across the sciences can be vastly diverse. Researchers use multiple types of file formats to record data in their studies. Some scientific data is numerical, while others may be audio survey recordings or videos of animal behaviours. To analyse this data, researchers use different applications and software. In some cases, they use applications such as the Microsoft Software Suite, but in others, they write Python or R code and use existing packages. Upon completing a study, these data, documents and code resources are often released online for verification, reuse and reproducibility purposes. Computational reproducibility (here used interchangeably with replication) is defined as the ability to obtain reported research results by re-executing original computational steps (National Academies of Sciences, Engineering, and Medicine, 2019). To maximize the opportunities for efficient discovery and reuse of these research outputs, they are deposited on the dissemination infrastructures that provide visibility on the web. Therefore, there is a demand for research dissemination infrastructures to support diverse datasets and make them FAIR.

The FAIR principles have emerged as guidelines to facilitate making digital research resources findable, accessible, interoperable, and reusable (Wilkinson et al., 2016). They have been widely recognized and adopted as the vision for research infrastructures supporting effortless data reuse. In practice, this is achieved with metadata, which can be described as data about data or adequate and machine-actionable documentation of the shared resources (Jacobsen et al., 2020). The first principle, findable, implies that research data should be described with metadata and should have a persistent identifier on the web. This metadata record should be shared in a data repository, which will facilitate its discoverability on the web. Accessible implies that the data should be shared via standard access protocols (but it does not mean that the data itself needs to be open access). Interoperable implies that file formats should be standard and that a description of data elements is available. Reusable means that the data should be assigned a license and usage rights and that its provenance is known, meaning that a re-user can understand what is in the data and how it was created, which is critical for its reuse. The FAIR principles stress the high importance of metadata, which led to the development of a number of detailed, standardized and community-used metadata schemas for research data.

In practice, however, research practices are constantly evolving, with the use of ever-increasing data volume, computing power and the complexity of computational methods and components. The shared scientific resources are becoming more diverse, including code, configuration files, workflows and containers, which creates new challenges in metadata implementation and its support in a research repository. To address these challenges, the data curation and dissemination community has ongoing conversations and developments of the metadata schemas. The study described in this paper is conducted for the purposes of informing the development of metadata for these conversations with a special focus on the Dataverse repository network.

The Dataverse Project provides an open-source data repository platform for sharing, archiving, and citing research data.[1] There are currently 80 installations of Dataverse data repositories around the globe, supporting institutional or national research and working together as a community to address the existing challenges. The Harvard Dataverse is the oldest and the largest installation in the repository network.[2]

This paper presents an exploratory data analysis of the open research datasets published on the Harvard Dataverse repository. The sample for the analysis is a description of over 40,000 open research datasets containing over 500,000 files. The analysis goal is to determine if and how research datasets cluster based on what researchers organically deposit together with the use of machine learning algorithms. Identifying a number of discrete groupings in the sample could inform future metadata development, which would be particularly useful in the Dataverse repository network.

[1] Dataverse Project: https://dataverse.org/

[2] Harvard Dataverse: https://data.harvard.edu/dataverse

The key problems I am aiming to address are the following:

- Shared datasets may include a number of different files, each of which may need to be supported differently. A quintessential example is a dataset that contains both data and code, which thus require different technical support for their dissemination, including licensing, attribution and storage.
- Many dissemination infrastructures do not support metadata for computational components (i.e., software, container and workflow files) necessary for reproducibility.

The paper's findings and contributions can be summarised as follows:

1. It provides evidence showing that the presence of research code has increased over the years in the research data publications.
2. It shows that open research data clusters in single-type file groups and an aggregate type containing various file types.
3. It proposes the use of a *Replication resource* metadata type, or in general, flexible metadata, which can bundle different types of objects such as *Datasets*, and *Software*.

The study is intended to facilitate open data sharing and reuse by providing insight into research data clusters, which could be valuable for metadata developments. It should be of interest to researchers, digital libraries and research infrastructure communities across the sciences.

Methods & Results

For the purposes of the analysis, I analyze a data sample describing the publicly shared research datasets from Harvard Dataverse. These are not the hundreds of datasets themselves, but only their metadata, which is stored on the repository for every deposited dataset. The unique persistent identifier of a dataset, a list of file types it contains, and the publication year are what comprise the studied data sample.

Each dataset contains files that are identified with a media type (also known as a mime type), which is a two-part label used to identify file formats on the web. On Dataverse repositories, these labels are used to enable file handling, like its preview in the browser. There are over 300 different media types identified on Harvard Dataverse. By mapping each media type into an object (Data, Text, Code, Document, etc.), we can determine the content of any dataset. I use nine different types of objects (archive, audio, code, data, document, image, shape, text, video) to group the files from the initial sample. An example of the mapping between mime types and objects is shown in Table 1. The original sample contained a list of media types for each dataset, which were then transformed into a list of objects for each dataset.

Table 1. The mapping of media (mime) types in the format "type/subtype" and objects (facets).

Media type	Object
`application/pdf`	Document
`application/zip`	Archive
`text/plain`	Text
`text/x-stata-syntax`	Code
`text/x-python`	Code
`text/tab-separated-values`	Data
`image/jpeg`	Image

Initially, the sample contained the content of about 45,000 datasets, but after a data cleaning step where the media types were mapped into objects, the sample counted 40,634 dataset entries with 586,169 files (fully mapped as objects). The size reduction occurred because many of the mime types that appeared a small number of times in the sample (many once or twice) were obscure and would require significant effort to be mapped into objects. Therefore, all datasets that had obscure mime types were removed from the analysis. After the cleaning step, the sample contains a DOI as a unique persistent identifier for each dataset and a count of each object from that dataset, formatted as shown in Table 2.

Table 2. Clean analysis data – each dataset (identified with a DOI) is assigned its content and their count.

DOI	Code	Data	Document	Img.	Text
10.7910/DVN/007GT	0	5	0	0	0
10.7910/DVN/00CIUU	11	9	0	0	0
10.7910/DVN/00IT1L	6	3	1	0	5
10.7910/DVN/00KDYS	0	9	1	0	0
10.7910/DVN/00ROYZ	0	7	0	0	0

We can immediately see that most datasets contain a small number of objects (Figure 1). Further, we see that over the years, the portion of datasets that contain code has been increasing (Figure 2.) at the Harvard Dataverse repository. The small number of objects per dataset may suggest that only a few object types are frequently shared together, such as data and code or data and documentation, indicating that we may see discrete clusters.

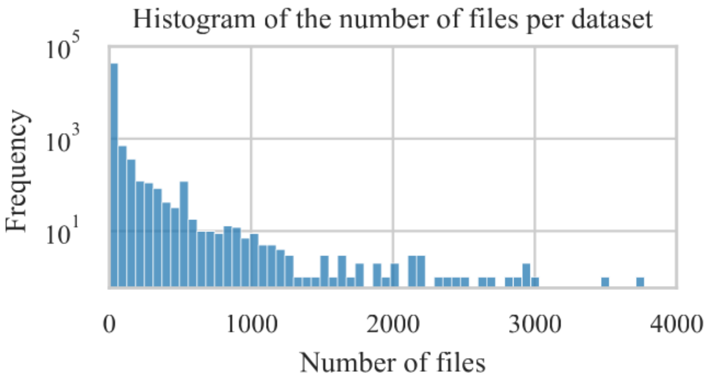

Figure 1. Histogram of object (file) counts per dataset.

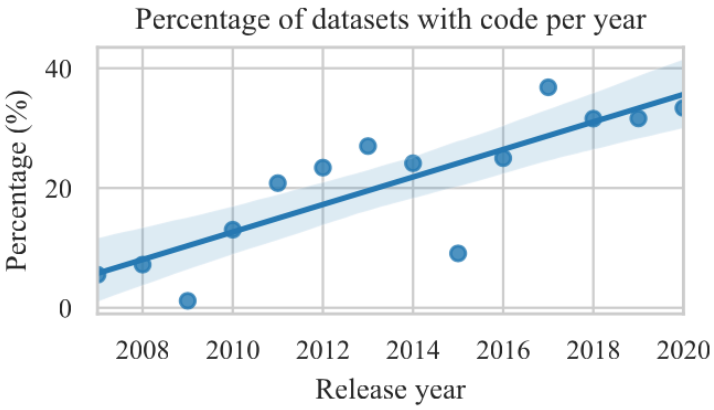

Figure 2. Percentage of datasets published at Harvard Dataverse repository that contain code files per year. Sample size: 40 634 datasets.

Clustering is a process where natural groupings within a set are determined, such that the items in each group exhibit more similarity to one another than to items in other groups. This analysis seeks to find if research datasets can be grouped according to their content and if so, identify meaningful structures that can be used to inform future developments in research infrastructure. I use techniques collected by Adolfssona et al. (2019) to assess the potential of the datasets in my sample to form clusters. I use the following approaches and algorithms:

1. Naïve approach;
2. Hopkins statistics;
3. Multimodality tests;
4. Visual assessment of (cluster) tendency (VAT) (Wang et al., 2010)
5. Principal component analysis (PCA)

Naïve cluster identification

A naïve approach identifies dataset clusters by counting how many times each combination of objects appears within the studied sample. Considering that the dataset contains nine types of objects, there are 2^9 - 1 = 511 possible dataset structures (-1 excludes the empty set). The sample contains a total of 140 object combinations (out of 511 possible). The most and least frequent combinations are shown in Figure 3.

We can observe that the twelve most frequent object clusters contain one or two object types and account for 70% of the whole sample. Further, we can see that code objects are often clustered together with data, documentation, and text objects (15% of the sample). There are 27 object combinations that appear only once in the sample. The naïve approach suggests that there are up to 140 possible object combinations, in different proportions, in the sample. In further analysis, we look for evidence to aggregate them into meaningful clusters.

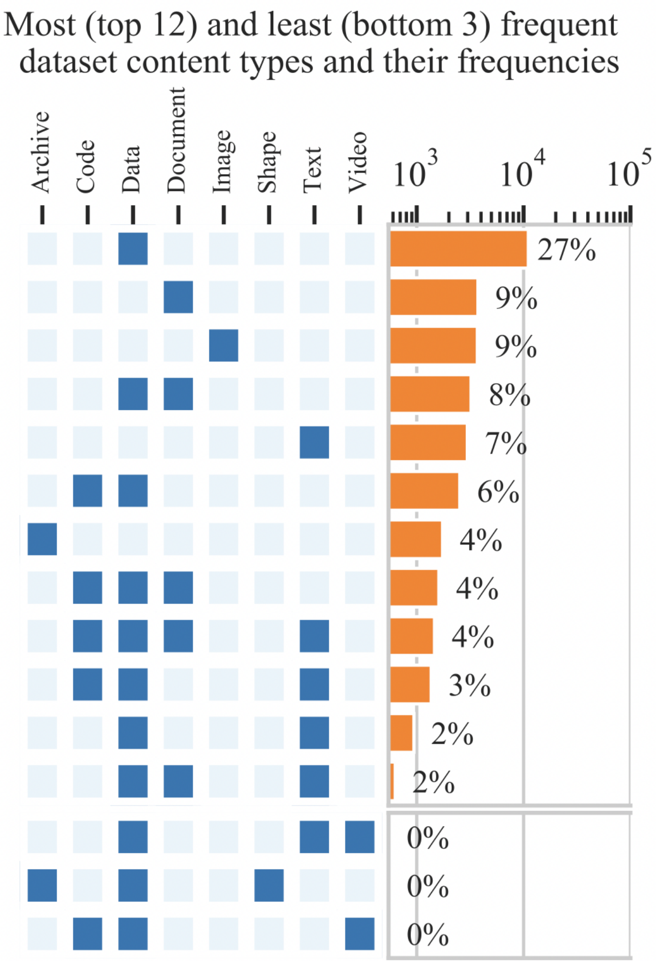

Figure 3. Most frequent dataset content types and their frequencies with a percentage of the total dataset. The most common dataset (27%) contains only data and is first on the list.

Hopkins statistic

The Hopkins statistic (Lawson & Jurs, 1990) is a statistical test that evaluates the clustering tendency of a dataset. It measures the spatial randomness of the data by comparing it to generated sample with uniform distribution and returns the probability that the data has a uniform random distribution. The Hopkins statistic score is between 0 and 1, where values close to 0 can be interpreted as high cluster tendency, and values above 0.3 express no clusterability. We obtain a score of h = 0.0026 for the full sample, and a score of h = 0.0064 when all single-type datasets are removed, both suggesting that the dataset is highly clusterable. However, Hopkins statistic is primarily a test against uniformity, meaning that data can be non-uniform but not suitable for clustering as the test does not ensure there is more than one cluster.

Multimodality tests

If a dataset contains multiple clusters, then there should be some identifiable separation between them. In particular, a histogram of pairwise distances should show a group of small distances within each cluster and large distances between the clusters. In case the dataset is homogeneous, it will not show such visual separation. When data is generated from a single bivariate normal distribution, it forms one cluster, and its pairwise distance and first principal component distributions are unimodal. By contrast, when the data is generated from multiple clusters, the pairwise distances and first principal component distributions are multimodal. Figure 4 has six modes suggesting clusterability. However, again it is important to note that the modes of data need not correspond with the actual number of clusters.

Multimodality statistical tests can formally determine if the set of distances has multiple modes, indicating multiple clusters. The dip test (Hartigan & Hartigan, 1985) is a widely used multimodality test that computes a statistic called the dip, which is defined as the maximum distance between the empirical distribution and the closest uniform distribution. It returns a *p*-value as the probability of observing the input being generated from a unimodal distribution (its null hypothesis). If only a single mode is present, the *p*-value will be large, suggesting that the data cannot be clustered. A small *p*-value, such as the one we observe here $p >= 0.001$ (the dip value of 0.12) suggests that multiple modes (and multiple clusters) are present.

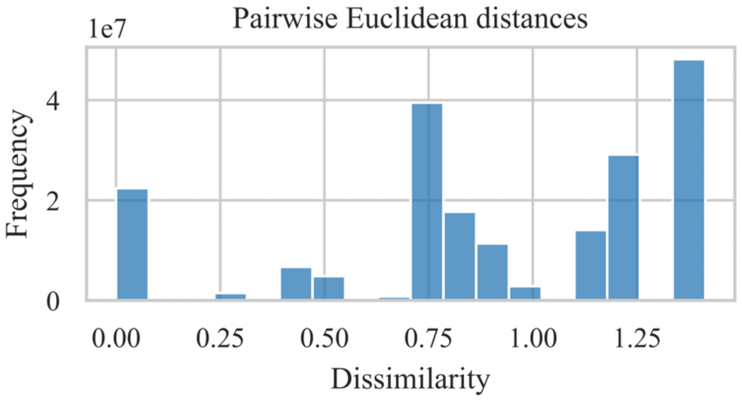

Figure 4. Pairwise Euclidean distances. Multiple modes in the distance distribution suggest the presence of multiple clusters.

Visual assessment of (cluster) tendency (VAT)

Visual assessment of tendency (VAT) (Bezdek & Hathaway, 2002) is an algorithm used for visually assessing the clustering tendency of multi-dimensional data. VAT creates a pairwise dissimilarity matrix of the data sample, and represents it as an *n x n* image. Its rows are reordered to reveal cluster structures as dark blocks along the diagonal of the image. Such visualization can be useful for obtaining insight into the number of clusters and their hierarchies. We can observe that the VAT image (shown in Figure 5, on the left) suggests that there are seven or more clusters in our sample, though some of the blocks are not visually clear.

VAT is effective in datasets that contain well-separated clusters, which may not be the case here. Its modification has been proposed for data with irregular structures (Wang et al., 2010). Improved VAT (iVAT) produces images of higher precision and should clearly show the number of clusters and their approximate sizes within the dataset. From Figure 5 (on the right), we are able to see four clear clusters (in black) and a cluster containing an indefinite number of clusters (shades of grey). The VAT tests do not seem to provide a conclusive result.

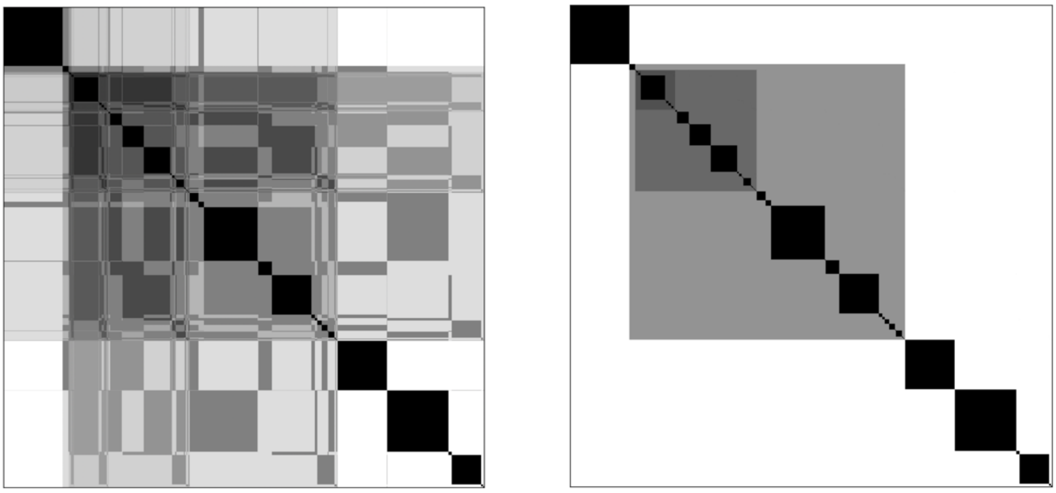

Figure 5. Visual assessment of (cluster) tendency (VAT) on the left and Improved VAT (iVAT) on the right, on a subsample of 5,000 records.

Principal component analysis (PCA)

Principal component analysis (PCA) is an unsupervised learning approach used in exploratory analyses that reduces data from high dimensions to lower dimensions while preserving the covariance in the data. The dimensionality reduction can help in visualizing high-dimensional datasets and intuitively judging whether they have meaningful clusters. If the data contains clusters, they should be spread out and visible when plotted out in a two-dimensional diagram when the PCA algorithm is applied.

For PCA, the sample is used in three different versions; the first one is the full sample, the second is the reduced sample where single-type datasets have been removed, and the third is a "binary" sample where each object's presence is marked with a 1 or 0. The scatter plots of the three sample versions after a PCA analysis in two dimensions do not show identifiable clusters (Figure 6). The explained variance ratio is the percentage of covariance explained by the reduced dataset. In this case, they are 51%, 61%, and 66%. This means that the sample variance is not fully preserved in two dimensions. The sample is used in PCA and plotted for a higher number of dimensions (up to five), but no clusters were observed.

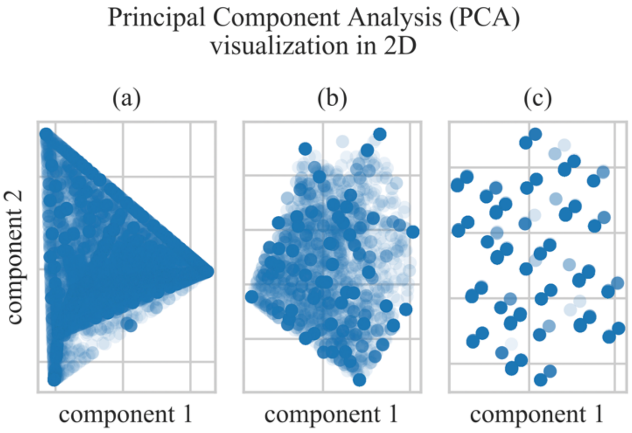

Figure 6. PCA results are obtained from the (a) full sample, (b) a sample without single-type records, and (c) binary sample (where each object presence is labelled with 1 or 0).

Discussion and Implications

The results of the statistical tests and the clustering algorithms described in the previous section are somewhat paradoxical. On one hand side, they suggest that the data is highly clusterable, but on the other, it seems that the sample cannot be cleanly clustered. Such a result can be interpreted in the following way: the sample does contain discrete clusters, but those are exclusively (or almost exclusively) single-type clusters (65% of the sample). The rest of the sample forms a cluster that cannot be meaningfully clustered any further.

Figure 7 shows the share of object types in the sample that supports this result interpretation. In other words, most datasets can be classified as a single-type object, but for the rest, an aggregate object type that incorporates various research objects is needed.

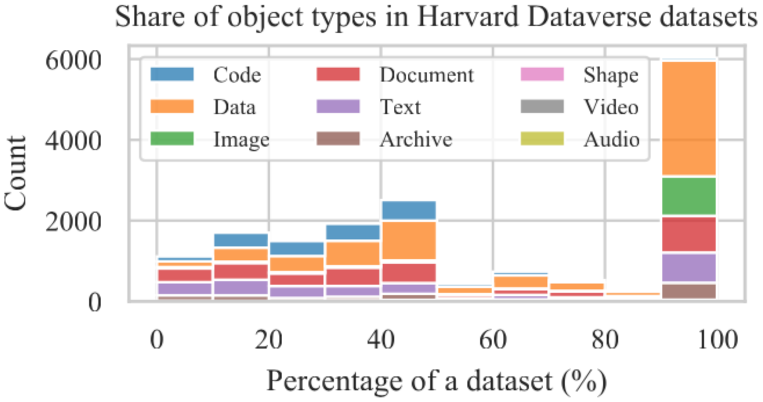

Figure 7. Share of object types in Harvard Dataverse datasets (percentage)

Metadata for research data

DataCite is a leading international organization that aims to improve and standardize data citation by enabling easier access to research data on the web, its recognition as research output and support for data archiving and reuse.[3] One of the significant roles of DataCite is the assignment of persistent identifiers and, in particular, digital object identifiers (DOI) to shared datasets belonging to its member research repositories and registries. The use of DOIs has been particularly notable in improving dataset accessibility, as they are typically displayed as a linkable URL which is immediately resolved to point to the dataset. DataCite also develops the DataCite Metadata Schema with key metadata properties necessary for consistent identification of a resource and its reuse.

The latest version of the DataCite Metadata Schema currently recognizes the following dataset types (*resourceTypeGeneral*): *Audiovisual, Book, BookChapter, Collection, ComputationalNotebook, ConferencePaper, ConferenceProceeding, DataPaper, Dataset, Dissertation, Event, Image, InteractiveResource, Journal, JournalArticle, Model, OutputManagementPlan, PeerReview, PhysicalObject, Preprint, Report, Service, Software, Sound, Standard, Text, Workflow, Other.*

Considering the resource type metadata, as used by DataCite, and with the idea to further interpret the analysis results, I transform the sample by grouping all datasets as either a single-type resource or an aggregate resource, in the following way:

[3] DataCite: https://datacite.org/

Objects	**Resource Type**	**% in the sample (count)**
Data, Archive, Shapefile	*Dataset*	37% (14859)
Audio, Video	*Audiovisual*	0.1% (73)
Code	*Software*	1% (378)
Document	*Report, Preprint, Journal Article, Dissertation etc.*	11% (4296)
Image	*Image*	10% (4101)
Text	*Text*	7% (2977)
Aggregate - with software code - without software code		34% 20% (8145) 14% (5805)

To illustrate the transformation of the sample, consider the following example. Code files are among the most fragile (software- and system-dependent) research artefacts and are often shared according to specific guidelines. Because of that, we can see single-object datasets or code objects grouped with text objects (as their documentation). Datasets with data objects or code objects (alone or accompanied by text) are grouped as a single type, i.e., *Dataset* or *Software*, respectively.

Given that the aims of the paper address the reuse of research resources in a data repository, the results suggest that a flexible metadata format may be an optimal solution. The aggregate cluster could be described with a resource type *Collection*, as it implies that the object contains various elements, though it may be a vague description for many of the datasets. It is important to note that a significant portion of the aggregate cluster contains datasets that contain research software or code. Specifically, that is about 60% of the aggregate cluster or about 20% of the full sample.

Replication metadata

Harvard Dataverse collaborates with many of the leading journals (Trisovic et al., 2022). Many datasets containing code and other objects represent replication packages meant to facilitate research verification and reproducibility of results published in journals. I argue that it would be beneficial for the research community to incorporate such a resource type to describe replication data that would be natively supported in research dissemination infrastructures. It would be particularly valuable for interdisciplinary or general-purpose repositories and registries.

The finding that most datasets contain single-type files has already been somewhat recognized in the communities developing research infrastructures. Many projects and platforms have emerged to address challenges for a single research output independently. In particular, every output is assigned specific metadata and would be published in a single-purpose repository such as:

1. Data in a data repository (i.e., Harvard Dataverse[4]),
2. Software in a software repository (i.e., GitHub[5]),

[4] Harvard Dataverse: https://dataverse.harvard.edu/

[5] GitHub: https://github.com/

3. Workflows at a workflow repository (i.e., WorkflowHub[6]),
4. Manuscripts at preprint repositories (i.e., ArXiv[7]), and so on.

Such an approach may result in scattered reporting and likely hinders research reproducibility. Reproducibility is fragile, as even incorrect file paths may result in being unable to rerun research code (Trisovic et al., 2022). Unless all these repositories are internally cross-referenced, a user may not be able to obtain all necessary research artefacts and reuse them. Therefore, it is more beneficial to enable support for diverse file formats in a single research infrastructure.

With *Replication resource* type, each published study would potentially be much easier to grasp and reproduce. All its resources would be gathered in a single bundle, including data, code, documentation, slides, and review.

Research Object (RO) is a metadata framework for capturing resources into citable reproducible packages, and represent an implementation of a *Replication resource* metadata. It uses standardized metadata based on schema.org to make these packages FAIR.[8] Further, RO-Crate acts as a collection of references to digital and physical objects, in any format, as a file or a URL, in a single linked-data metadata document (Soiland-Reyes et al., 2021, Sefton et al., 2021). As such, it provides an integrated view of research resources that can be used for reproducing and reusing existing studies. RO-Crate enables exceptional flexibility in creating a replication package. Essentially, any type of file can be deposited together, and each will have machine-readable metadata with its contextual information to aid in decision-making when reusing the package.

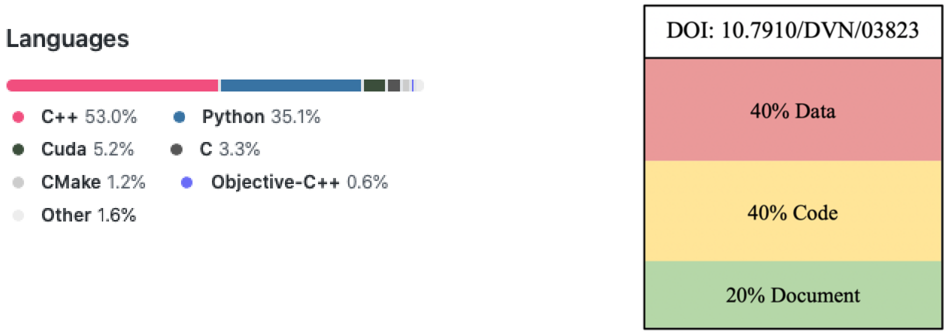

Figure 8. Visual representation of aggregate datasets.

Finally, research datasets could be made more seamlessly reusable by visualizing their content. That way, users could intuitively grasp its utility for their use case. GitHub is the software repository hosting platform used for software development, collaborative work, version control with git and software dissemination. It is currently the largest source code host and is commonly used for open-source projects. It incorporates a high degree of flexibility in its software deposits. For instance, while its users may deposit any file type, special attention is taken to the source code and its programming language. Figure 8 (left) shows a visual representation of a programming language share in a single repository. Figure 8 (right) shows a visual representation proposal, inspired by the analysis results and the GitHub interface, that research data repositories could employ in their UI.

[6] WfCommons: https://wfcommons.org/
[7] ArXiv: https://arxiv.org/
[8] Schema.org: https://schema.org/

Acknowledgements

A. Trisovic is funded by the Alfred P. Sloan Foundation (grant number P-2020-13988). The author thanks the Dataverse team for their help in retrieving data for this analysis and the Harvard Data Commons team for useful discussions on related topics.

References

Adolfsson, A., Ackerman, M., & Brownstein, N. C. (2019). To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*, 88, 13–26. https://doi.org/10.1016/j.patcog.2018.10.026.

Bezdek, J. C., & Hathaway, R. J. (2002). VAT: a tool for visual assessment of (cluster) tendency. *Proceedings of the 2002 International Joint Conference on Neural Networks*. IJCNN'02 (Cat. No.02CH37290). https://doi.org/10.1109/ijcnn.2002.1007487.

DataCite Metadata Working Group (2021). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.4. DataCite e.V. https://doi.org/10.14454/3w3z-sa82.

Hartigan, J.A., & Hartigan, P.M. (1985). The Dip Test of Unimodality. *Annals of Statistics*, 13 (1), 70-84. https://doi.org/10.1214/aos/1176346577.

Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., Goble, C., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R. W. W., Imming, M., Jeffery, K. G., … Schultes, E. (2020). FAIR Principles: Interpretations and Implementation Considerations. *Data Intelligence*, 2 (1–2) , 10–29. https://doi.org/10.1162/dint_r_00024.

Lawson, R. G., & Jurs, P. C. (1990). New index for clustering tendency and its application to chemical problems. *Journal of Chemical Information and Computer Sciences,* 30 (1), 36–41. https://doi.org/10.1021/ci00065a010.

National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303.

Sefton, P., Carragáin, E. Ó., Soiland-Reyes, S., Corcho, O., Garijo, D., Palma, R., ... & Portier, M. (2021). RO-Crate Metadata Specification. Retrieved from: https://www.researchobject.org/ro-crate/.

Soiland-Reyes, S., Sefton, P., Crosas, M., Castro, L. J., Coppens, F., Fernández, J. M., Garijo, D., Grüning, B., La Rosa, M., Leo, S., Carragáin, E. Ó., Portier, M., Trisovic, A., Community, R.-C., Groth, P., & Goble, C. (2021). Packaging research artefacts with RO-Crate. https://doi.org/10.48550/ARXIV.2108.06503.

Trisovic, A., Lau, M. K., Pasquier, T., & Crosas, M. (2022). A large-scale study on research code quality and execution. *Scientific Data*, 9 (1). https://doi.org/10.1038/s41597-022-01143-6.

Wang, L., Nguyen, U. T. V., Bezdek, J. C., Leckie, C. A., & Ramamohanarao, K. (2010). iVAT and aVAT: Enhanced Visual Analysis for Cluster Tendency Assessment. *Advances in Knowledge Discovery and Data Mining*, 16–27. https://doi.org/10.1007/978-3-642-13657-3_5.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016) The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3 (1), 1–9. https://doi.org/10.1038/sdata.2016.18.